M. PETRUSZEWYCZ

Contribution pour servir à l'étude du choix que fit A. A. Markov d'un domaine d'application de sa théorie des chaînes

Mathématiques et sciences humaines, tome 66 (1979), p. 43-49.
<http://www.numdam.org/item?id=MSH_1979__66__43_0>

Math. Sci. hum. (17e année, n°66, *1979*, p.43-49)

CONTRIBUTION POUR SERVIR A L'ETUDE DU CHOIX QUE FIT A.A. MARKOV D'UN DOMAINE D'APPLICATION DE SA THEORIE DES CHAINES

M. PETRUSZEWYCZ

C'est une citation très brève faite par Maistrov [25] qui a attiré mon attention sur un article de V. Ja. Bunjakovskij (1804-1889). Il a paru intéressant d'en donner un résumé très contracté et d'en citer presque textuellement la fin car peut-être est-on ici en présence de l'une des origines possibles du domaine d'application choisi par Markov pour illustrer sa théorie des chaînes.

Ce long article (p.36 à 49) a paru dans le numéro 3 de 1847 de *Sovremennik* (Le Contemporain) revue littéraire fondée par Puškin. A cette date elle a pour "directeur idéologique" le réputé critique V. Belinski et au nombre de ses rédacteurs, le poète N. Nekrassov qui lui donna à partir de 1846 un ton démocrate. Continueront à y paraître les romans et nouvelles des romanciers russes universellement connus, mais aussi des enquêtes d'information économique et sociologique, des essais, des critiques d'ouvrages philosophiques et économiques. On trouvera ci-dessous un résumé de l'article de Bunjakovskij qui le premier a écrit en langue russe un traité de Calcul des Probabilités, paru en 1846 :
Osnovnaja matematiceskoj teorij verojatnostej : Base de la théorie mathématique des probabilités (1).

(1) A la demande de Melle M. Guy, Conservateur à la Bibliothèque du Grand Palais, la Bibliothèque Lénine m'a obligeamment adressé un microfilm de cet ouvrage, mais il semble qu'il n'y ait rien se rapportant au problème étudié. L'auteur tient cependant à remercier la personne et les organismes ci-dessus cités, la lecture de cet ouvrage étant fort intéressante.

L'article dont je vais parler a pour titre :

О Возможности введения определительных мере доверия к результатам некоторых наук наблюдательных и преимущественно статистики.

O vozmoznosti vvedenija opredelitel'nyh mere doverija k rezul'tatam nekotoryh nauk nabljudatel'nyh i preimuscestvenno statistiki.

Sur la possibilité de l'introduction de mesures définies de confiance dans les résultats de quelques sciences d'observations, principalement la statistique.

Bunjakovskij souligne pour commencer le rôle exemplaire que joue l'astronomie pour les sciences d'observations grâce à un remarquable degré de précision dans les résultats obtenus à partir d'un grand nombre de *résultats numériques* (1). En présence de nombreuses mesures il est usuel et commode d'utiliser la *moyenne arithmétique* mais cet instrument n'aura de valeur qu'autant qu'on pourra l'assortir d'un "intervalle de confiance" : (мера доверия mera doverija). Il donne l'exemple de *six* mesures de distance dont la moyenne est 20 sagènes 1 archine 1 verchok (2). Cette valeur n'est sûrement pas exacte : le problème est de "savoir de combien elle est plus grande ou plus petite que la valeur "précise" (точное = točnoe = précis)". Il est évident que ce problème n'a pas de solution sinon il n'y aurait pas de problème du tout. On est donc amené à le reformuler ainsi : quelle confiance peut-on avoir dans ce résultat ? Et il définit le "degré de confiance (степень доверия : stepen' doverija)" en un quelconque résultat numérique comme la "*probabilité* (вероятность = verojanost') au sens scientifique". Bunjakovskij décrit alors une urne contenant 1000 boules : 999 noires et 1 blanche ; si on fait un tirage au hasard, présupposant qu'aucune boule n'a plus de chance de sortir qu'une autre (textuellement "tomber dans la main : попасть в руку popast' v ruku"),la confiance dans la supposition que la boule sortie sera noire est 999 fois plus grande que la mesure de l'espérance dans l'événement contraire : apparition de la boule blanche. On peut prendre pour mesure de confiance de l'apparition d'1 boule noire un nombre arbitraire et l'autre 999 fois plus petit, ces deux nombres servant de mesure de

---

(1) Les italiques figurent dans le texte russe.

(2) Anciennes unités de mesure : 1 sagène = 2,13m ; 1 archine = 0,711m ; 1 verchok = 0,44m.

confiance pour les deux événements complémentaires. Arbitrairement les mathématiciens ont décidé que la somme de ces deux nombres devait être égale à l'*unité*, mesure de la certitude. Les fractions ainsi obtenues 999/1000 et 1/1000 s'appellent des probabilités. L'auteur donne alors la définition de *la probabilité d'un événement quelconque* comme *la fraction dont le numérateur est le nombre de cas favorables à cet événement et le dénominateur le nombre de tous les cas possibles*. Il revient alors sur les conditions d'équiprobabilité qu'il appelle textuellement "*d'égale possibilité* равновозможных: ravnovozmojnyh ; adjectif qu'il utilise pour qualifier "tant les cas favorables (благоприятствующий : blagoprijatstvujuscij [à l'événement choisi] que les cas contraires неблагоприятствующий neblagoprijatstvujuscij ". Il ajoute que "dans les cas "d'inégale possibilité" неравновозможных neravnobozmojnyh) l'analyse des probabilités aboutit à des règles spéciales à l'aide desquelles on ramène le problème dans le cadre de l'hypothèse d'égale possibilité.

Le praticien des sciences d'observations, l'expérimentateur sait qu'au terme d'une série de mesures il devra donner pour résultat une *approximation* assortie du plus grand degré possible de confiance. Par des méthodes mathématiques rigoureuses on démontre qu'un nombre croissant de mesures peut, sous certaines conditions, atteindre le degré désiré de probabilité c'est-à-dire la moyenne des mesures ne sera différente de la "valeur vraie" (истинна : istinna : authentique, vraie) que d'une très petite quantité par rapport à cette valeur et sera d'autre part insensible à l'expérimentation même. Il donne un exemple : on mesure une distance 1000 fois et on veut que la valeur moyenne des mesures ne diffère de la valeur réelle " истинна : istinna ", que de 1/100. Sans expliciter davantage le modèle auquel il se réfère il écrit que si la distance cherchée est x et la moyenne des observations a la probabilité sera égale à 999/1000 (c'est-à-dire qu'on pourra parier (1) à 999 contre 1) que x se situe entre les limites $a-a/100$ et $a+a/100$. Si on veut resserrer les limites, par exemple les fixer à $a-a/500$ et $a+a/500$, par cela même on ne pourra avoir alors une probabilité aussi élevée. Suit une évocation de *la méthode des moindres carrés* et la citation sans explication de *la loi des grands nombres* si rapides qu'on peut se demander ce que cela a pu évoquer chez le lecteur habituel de la revue. Il arrive alors au but de l'article qui est de plaider que la statistique devrait s'inspirer

---

(1) L'auteur utilise une expression "derjat' zaklad" : précisément "mettre en gage" ; actuellement on utiliserait plus probablement "держать пари derjat' pari : parier".

de ces pratiques pour assurer ses résultats. Son domaine d'application privilégié est la démographie : Bunjakovskij a calculé pendant plusieurs années l'effectif du contingent incorporable pour l'armée russe. Il reproduit en détail les calculs de Laplace pour l'estimation de la population globale française à partir des relevés sur 3 ans dans 30 communes françaises des naissances des deux sexes et l'erreur qui s'attachait à cette estimation. Il plaide ensuite pour l'extension de ces méthodes de relevé de données et des résultats qu'on en peut tirer à l'étude du niveau d'instruction dans un pays donné, par exemple.

Il évoque enfin rapidement les travaux dont il s'inspire : *Condorcet* (Essai sur l'application de l'Analyse à la Probabilité des Décisions), *Poisson* (Recherches sur la probabilité des jugements). L'idée principale est la suivante : les résultats numériques et habituellement les *moyennes arithmétiques* sont assorties partout où cela est possible d'une *mesure de confiance*. "Le concept même de mesure de confiance peut s'exprimer au moyen de la probabilité que le résultat obtenu se situe entre les limites trouvées ainsi qu'il a été expliqué ci-dessus [ou autrement en se conformant aux règles habituelles du calcul des probabilités]. Les statisticiens eux-mêmes montreront que c'est la pratique qu'ils considèrent comme la plus commode".

"Peut-être objectera-t-on que le statisticien en se vouant presqu'exclusivement à sa science n'a plus le loisir de s'occuper de l'étude approfondie des résultats les plus profitables de la théorie des observations, l'une parmi les plus difficiles dans le domaine de l'analyse des probabilités. A ceci on peut répondre que pour l'observateur il n'y a aucune nécessité, comme pour le mathématicien, d'avoir une parfaite connaissance avec tous les raffinements analytiques de la déduction détaillée des formules de cette théorie. Le but est atteint lorsque l'observateur sait utiliser les formules démontrées, mais pour cela il faut un manuel d'utilisation des observations. L'élaboration d'un tel manuel didactique à la portée de tous pour l'utilisation 'la plus efficace' des démonstrations relatives aux observations, c'est au mathématicien connaissant les exigences de la statistique de l'assumer. En ce qui concerne la statistique, il est indispensable d'avoir des connaissances de mathématiques élémentaires pour ne pas être embarrassé par la traduction numérique des formules générales. En effet, indépendamment de l'analyse des observations n'arrive-t-il pas qu'il faille effectuer dans des questions de type financier des calculs exigeant des considérations variées sur les intérêts composés : alors on doit nécessairement se plier aux formules algébriques et utiliser des tables de logarithmes.

D'ailleurs dans le cas de statistiques douteuses on peut recourir au mathématicien qui doit avoir une connaissance parfaite du langage des nombres. Le mathématicien doit éviter l'erreur dans laquelle tombent les observateurs qui attribuent à des résultats numériques des pondérations égales (достоинство : dostoinstvo : qualité. L'auteur oppose равньая достоинства: ravn'aja dostoinstva à степени доверия (stepeni doverija) alors qu'au contraire ces résultats ne méritent pas un égal degré de confiance".

"L'idée de l'introduction d'indications plus définies dans les résultats numériques s'est présentée, vraisemblablement, plus d'une fois déjà aux statisticiens. On doit supposer qu'elle n'a pas été mise à exécution à ce jour en raison d'une part des difficultés particulières rencontrées dans son application, et d'autre part en raison de l'absence d'un manuel exposant en un langage accessible à tous les instruments de la difficile théorie des résultats les plus utiles. Quoi qu'il en soit il semblerait permis au mathématicien d'espérer que le projet d'une statistique précise soit dans une certaine mesure possible. On est près même de prédire que dans quelques temps, et peut-être très vite, la statistique, à la suite des sciences d'observations parviendra à une exactitude qui cédant en son essence aux résultats astronomiques sera par contre de beaucoup supérieure à celle du présent état de cette science.

Ici devrait se conclure l'article, mais, par analogie, qu'il me soit permis d'ajouter quelques mots au sujet d'une autre application du calcul des probabilités que, semble-t-il, personne n'a encore indiquée. La nouvelle application se rapporte aux recherches grammaticales et étymologiques ainsi qu'à la philologie comparative. Pour autant qu'au premier coup d'oeil de telles recherches semblent étrangères à l'analyse mathématique, cependant on peut dire avec assurance que sous ce rapport se présente un vaste champ pour des applications mathématiques rigoureuses. Mon affirmation ne se fonde pas sur des suppositions et des conjectures plus ou moins précaires mais sur un examen critique de la discipline, *sur quelques essais que j'ai déjà réalisés et sur des formules analytiques que j'ai déduites pour définir les probabilités numériques de diverses dérivations linguistiques* (1).

Ainsi la mesure de confiance pour n'importe quelle étymologie par exemple peut être déterminée approximativement par un *nombre*, et par son degré de

(1) Les italiques ne sont pas ici de l'auteur (sauf pour numérique).

proximité par rapport à l'unité ou à la certitude ce qui permet de juger de sa supposée authenticité. Ce n'est pas maintenant le lieu de rentrer dans des détails à ce sujet sur lequel j'ai seulement voulu, pour l'instant, attirer l'attention. Mais pour montrer directement de quelle façon de semblables recherches peuvent entrer dans le domaine de la mathématique appliquée, il ne sera pas superflu d'énumérer quelques indications numériques ou matériaux relevant de leur élaboration. Quand on traite d'un langage, on suppose avant tout que l'on en a une *description numérique* détaillée ou, comme nous l'appellerons sa *statistique* c'est-à-dire des indications numériques sur le nombre total des mots de cette langue, sur la distribution de ces mots selon les parties du discours, selon leur nombre de lettres, selon la lettre initiale, selon les terminaisons, etc... Ici même se classent les connaissances sur les règles générales, sur les exceptions de tous genres, sur les mots empruntés sans aucun doute à d'autres langues et ainsi de suite. Voilà des données numériques dont l'analyse rigoureuse exige sans aucun doute la considération des mathématiciens. En possession de semblables données pour deux ou plusieurs langues on peut les comparer sous différents rapports et les résultats obtenus revêtiront une certaine autorité que ne peuvent toujours justifier les philologues dans le présent état de la science.

Assurément l'élaboration de ce que j'ai appelé la *statistique du langage* est très fatigante et selon toute vraisemblance les philologues déclareront qu'un tel travail est presqu'inutile pour le motif que le gain présumé du côté de l'exactitude de la conclusion dans le domaine linguistique ne les dédommagera pas du temps perdu. Nous ne prendrons pas sur nous de résoudre la question du degré de justesse de cette affirmation.

Peut-être, si une autre occasion se présente, je publierai mes recherches théoriques sur le sujet que j'ai seulement évoquées ici. Quant à l'utilisation pratique des formules générales, ne disposant pas de données arithmétiques détaillées sur les langues, on devra se limiter à un petit nombre d'exemples. D'ailleurs, pour donner à un semblable travail le degré de complétude qui lui est dû quant aux données philologiques, il va de soi que le mathématicien doit absolument entrer en rapport avec les spécialistes de ce domaine qui lui est plus ou moins étranger".

A ce jour de Paris il n'a pas été possible de trouver une publication ultérieure de Bunjakovskij où il aurait présenté des statistiques

lexicales ou leur exploitation. Cependant on ne peut plus attribuer, semble-t-il, à la lecture du texte ci-dessus, le titre d'initiateur que j'attribuais en conclusion de l'Annexe IV de [45], à Morozov. Celui-ci reste quand même un pionnier. Mais plus d'un demi-siècle sépare les publications des deux savants et nous ne connaissons pas les maillons manquants de l'histoire de la statistique lexicale en langue russe.